ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΗΛΕΚΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΡΧΕΙΩΝ

1η άσκηση

Ημερομηνία παράδοσης: 31 Μαρτίου 2016
Η άσκηση είναι ατομική

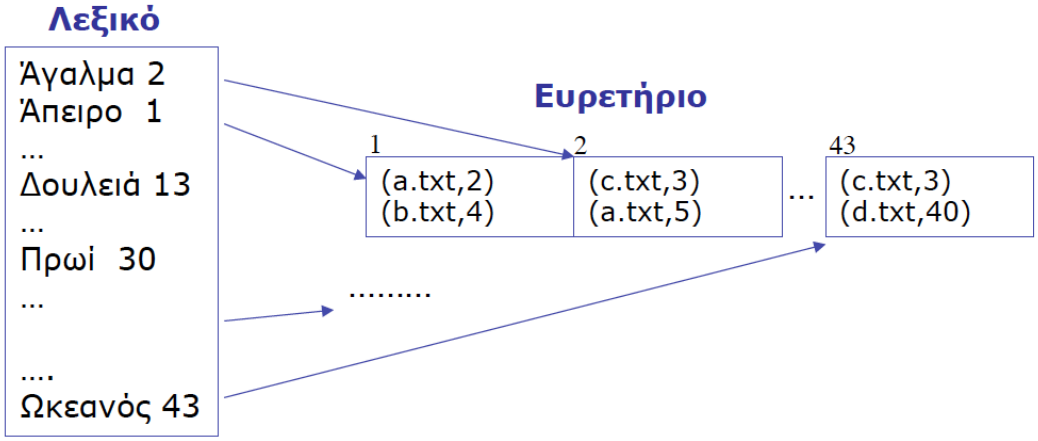
Επεξεργασία Αρχείων

Κατασκευάστε ένα πρόγραμμα που δέχεται ένα ή περισσότερα αρχεία κειμένου (όπως αυτά που σας δίνονται ενδεικτικά) και δημιουργεί μια δομή δεδομένων στο δίσκο που απαντά σε ερωτήσεις όπως «**βρες κείμενα που περιέχουν την λέξη 'άγαλμα'**» ή οποιαδήποτε άλλη λέξη.

Γράψτε ένα πρόγραμμα που δέχεται ένα ή περισσότερα αρχεία κειμένου (όπως αυτά που σας δίνονται) και κατασκευάζει την παρακάτω δομή αρχείου στο δίσκο. Η δομή αρχείου αποτελείται από το «**Λεξικό**» και το «**Ευρετήριο**». Η δομή αρχείου στο σχήμα δηλώνει ότι η λέξη «Άπειρο» του Λεξικού υπάρχει στην 1η σελίδα του Ευρετηρίου. Εκεί βρίσκεται η πληροφορία ότι ή λεξη υπάρχει στο αρχείο a.txt στην θέση 2bytes από την αρχή του αρχείου και στο αρχείο b.txt στην θέση 4bytes από την αρχή του αρχείου.

Κατασκευή της δομής δεδομένων στην Κεντρική Μνήμη (3 Μονάδες)

Αρχικά βρείτε όλες τις λέξεις διαβάζοντας όλα τα αρχεία εισόδου και κατασκευάστε τη δομή του σχήματος στην κεντρική μνήμη. Το Λεξικό είναι ένας πίνακας (array) με όλες τις λέξεις που υπάρχουν στα κείμενα (κάθε λέξη εμφανίζεται μια φορά). Ταξινομήστε τον πίνακα αλφαβητικά. Κάθε λέξη παριστάνεται από ένα string (αν θέλετε με μέγιστο μήκος 12 χαρακτήρων) και έναν ακέραιο αριθμό. Την τιμή του ακεραίου θα την αποφασίσετε όταν αντιγράψετε το δομή στο δίσκο. Κάθε λέξη δείχνει στο Ευρετήριο. Στην κεντρική μνήμη το Ευρετήριο ειναι μία λίστα.



Κατασκευή Δομής Αρχείου (4 μονάδες)

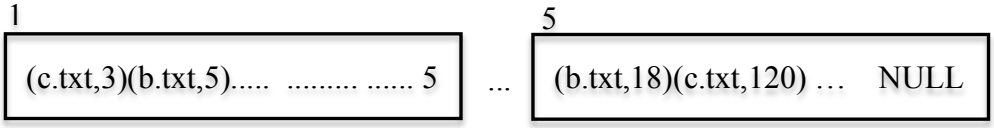
Αντιγράψτε την δομή δεδομένων στο δίσκο ακολουθώντας την παρακάτω διαδικασία.

Η σελίδα αρχείου έχει μέγεθος 128bytes. Φτιάξτε ένα buffer 128bytes στην κεντρική μνήμη και αρχίστε να το γεμίζετε με ζεύγη string – integer απο το Λεξικό. Όταν γεμίσει ο buffer (θα χωρέσει 9 τέτοια ζεύγη και θα περισσέψουν 2 bytes) γράψτε μία νέα σελίδα στο τέλος του αρχείου. Αδειάστε τον buffer και συνεχίστε το ίδιο μέχρι να αντιγραφεί το Λεξικό στο αρχείο.

Το Ευρετήριο είναι ένα αρχείο του οποίου κάθε σελίδα αποθηκεύει ζεύγη της μορφής (**όνομα αρχείου – θέσεις bytes από την αρχή του κειμένου)** π.χ. η λέξη «Άπειρο» στο σχήμα, συνοδεύεται στο Λεξικό από τον ακέραιο 2 που σημαίνει ότι δείχνει στην $2^{η}$ σελίδα του Ευρετηρίου. Η σελίδα 2 στο ευρετήριο έχει τις εγγραφές (b.txt,4) (a.txt,2) που δηλώνουν ότι η λέξη «Άπειρο» υπάρχει στο αρχείο "a.txt" στην θέση 2bytes από την αρχή του και στο αρχείο "b.txt" στη θέση 4bytes από την αρχή του. Αν μία λέξη υπάρχει στο Λεξικό τότε υπάρχει γι αυτήν τουλάχιστον μία σελίδα στο Ευρετήριο.

Για κάθε δήλωση π.χ. (c.txt,3) στο Ευρετήριο χρειάζονται 12bytes δηλαδή 8 για το όνομα του αρχείου και 4bytes για την θέση της λέξης στο αρχείο. Το Ευρετήριο αποτελείται από σελίδες μεγέθους 128bytes. Άρα κάθε σελίδα χωράει 10 εγγραφές και περισσεύουν 8bytes (4 από αυτά ίσως χρειαστούν για να συνδέσετε μεταξύ τους δύο σελίδες όπως περιγράφεται παρακάτω). Αντιγράψτε το Ευρετήριο από την κεντρική μνήμη στο δίσκο.

Αν μία λέξη επαναλαμβάνεται συχνά στα κείμενα τότε μπορεί να έχει πολλές εγγραφές στο Ευρετήριο που δεν χωράνε όλες σε μία σελίδα. Μόλις γεμίσει μια σελίδα στο Ευρετήριο, δημιουργείται μία νέα στο τέλος του αρχείου και συνδέεται με την πρώτη όπως στο σχήμα: η σελίδα 1 συνδέεται με την σελίδα 5 (μπορεί όμως να ειναι συνεχόμενες σελίδες στο αρχείο δηλ η 1 να δείχνει στην 2 κλπ).



Αναζήτηση στη δομή δεδομένων (3 Μονάδες)

Γράψτε ένα πρόγραμμα που στην είσοδό του δέχεται μία λέξη και ως απάντηση επιστρέφει όλα τα κείμενα στα οποία υπάρχει η λέξη και την θέση στην οποία βρίσκεται.

Αρχικά, γράψτε μια συνάρτηση που εκτελεί δυαδική αναζήτηση σε αρχείο και βρίσκει αν υπάρχει μία λέξη στο Λεξικό: διαβάσετε την μεσαία σελίδα του αρχείου φέρνοντάς την στην κεντρική μνήμη (κοστίζει μια πρόσβαση δίσκου). Κάντε αναζήτηση μέσα στην σελίδα που φέρατε (αυτό δεν κοστίζει σε προσβάσεις δίσκου). Αν την βρείτε, η συνάρτηση επιστρέφει το περιεχόμενο της θέσης δηλαδή το όνομα του αρχείου και τον

ακέραιο. Αν δεν την βρείτε τότε διαβάστε την μεσαία σελίδα από το αριστερό ή δεξιό μισό του αρχείου ανάλογα με το αν η λέξη που ψάχνετε είναι αλφαβητικά πριν ή μετά τις λέξεις που υπήρχαν στη μεσαία σελίδα του αρχείου (αν χρειαστεί να συνεχίσετε την ίδια διαδικασία).

Κάθε διάβασμα σελίδας κοστίζει μία πρόσβαση στο δίσκο. Από την κάθε σελίδα έχετε στο Ευρετήριο για τα αρχεία που περιέχουν την λέξη που ψάχνετε. Κάθε διάβασμα σελίδας στο Ευρετήριο κοστίζει επίσης μία πρόσβαση στο δίσκο. Το αποτέλεσμα που εμφανίζεται στην οθόνη έχει την μορφή

«κείμενο 'c.txt' περιέχει την λέξη 'άγαλμα' στην θέση 3»
«κείμενο 'a.txt' περιέχει την λέξη 'άγαλμα' στην θέση 5»
«κείμενο 'b.txt' περιέχει την λέξη 'άγαλμα' στην θέση 18»
«κείμενο 'c.txt' περιέχει την λέξη 'άγαλμα' στην θέση 120»

Για κάθε αναζήτηση επιστρέφεται επίσης ο αριθμός **κ** που δείχνει πόσες προσβάσεις στο δίσκο κόστισε η αναζήτηση (ο αριθμός προσβάσεων δίσκου στο Λεξικό προστίθεται στο αριθμό προσβάσεων δίσκου που έγιναν στο Ευρετήριο).

Παραδοτέα: Ένα συμπιεσμένο zip αρχείο που περιέχει ότι ζητείται παρακάτω:
Α) ο κώδικας περιέχει σχόλια που εξηγούν την υλοποίηση
Β) μία έκθεση που περιγράφει σε 1-2 σελίδες πως φτιάχτηκε ο κώδικας (δηλ. για κάθε ερώτημα ποια είναι η γενική ιδέα της λύσης σε 3-4 προτάσεις), υπάρχουν σαφείς οδηγίες μετάφρασης από compiler και εκτέλεσης, τι λάθη έχει (αν έχει, περιπτώσεις που δεν δουλεύει το πρόγραμμα, ή περιπτώσεις που κάνει περισσότερα από όσα σας ζητεί η άσκηση, τι χρησιμοποιήσατε από έτοιμα προγράμματα ή πηγές πληροφόρησης.
Υποδείξτε ακόμα και πηγές στο www όπως Wikipedia ή ακόμα και συναδέλφους που σας βοήθησαν στην άσκηση. Αυτό θα βοηθήσει να μη θεωρηθεί ως αντιγραφή μια άσκηση που μοιάζει με μία άλλη ή μια πηγή στο www. Όμως, αν μία άσκηση είναι αντιγραφή θα μηδενιστεί.
Οι ασκήσεις βαθμολογούνται με άριστα εφόσον
Α) το zip είναι πλήρες
Β) Οι κώδικες περνούν από compiler και εκτελούνται κανονικά και σωστά σε windows ή Linux περιβάλλον